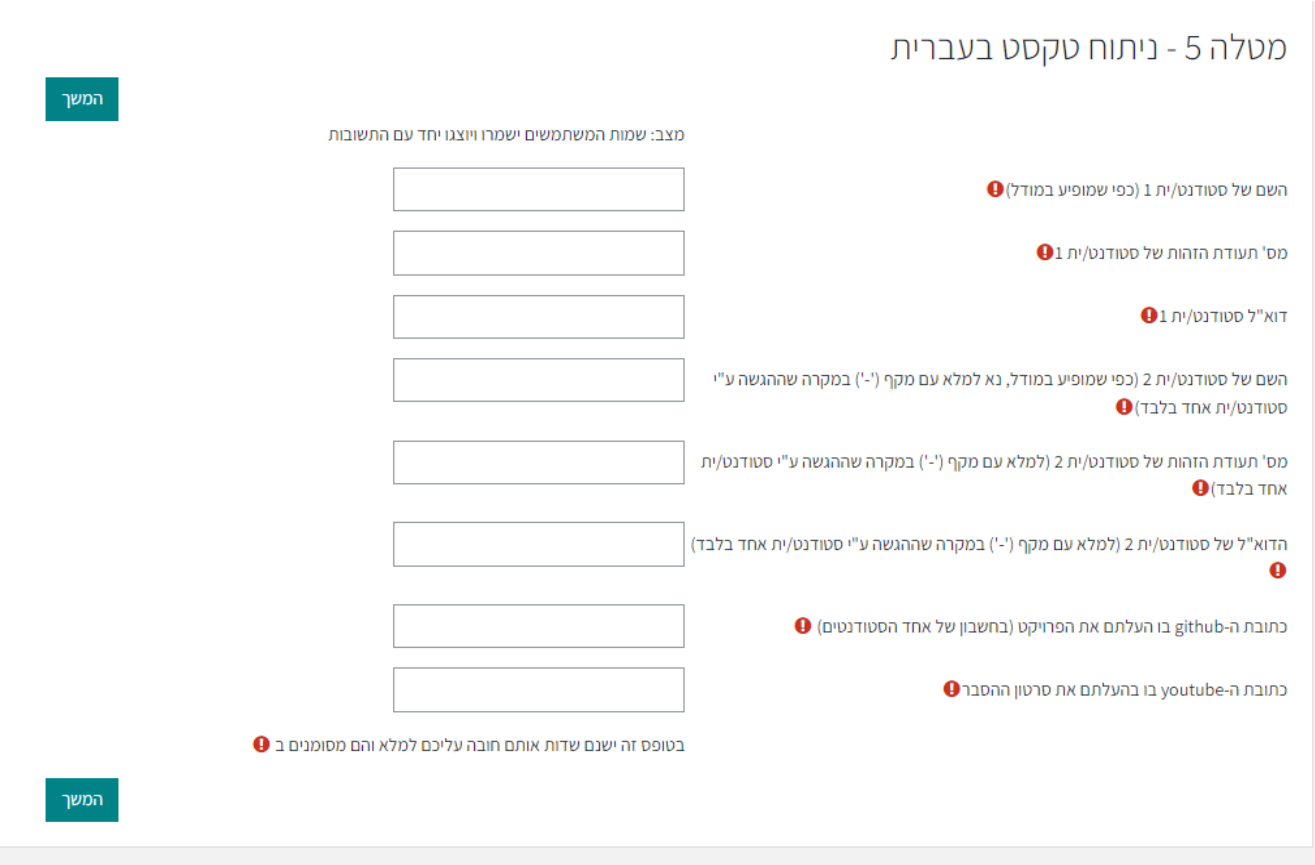
למידת מכונה – מטלה 5 - ניתוח טקסט בעברית – מסמך הסבר

תאריך הגשת המטלה

את המטלה יש להגיש עד יום ראשון בערב ה-14/5. הגשה באיחור עד ה- 18/5

אופן ההגשה

- ניתן להגיש את העבודה ביחידים או בזוגות.
- יש לשלוח בטופס את שמות המשתתפים, מס' ת.ז. ומייל של כל משתתף
- יש להעלות את העבודה לדף ה-github של אחד המשתתפים ולשלוח לנו קישור (בהגשה)
- יש להקליט סרטון ב-YouTube הסביר את הקוד של המשתתפים ולהעביר לנו את הקישור


מטלה 5 - ניתוח טקסט בעברית

המשך

מצב: שמות המשתמשים ישמרו ויוצגו יחד עם התשובות

השם של סטודנט/ית 1 (כפי שמופיע במודל)

מס' תעודת הזהות של סטודנט/ית 1

דוא"ל סטודנט/ית 1

השם של סטודנט/ית 2 (כפי שמופיע במודל, נא למלא עם מקף ('-') במקרה שההגשה ע"י סטודנט/ית אחד בלבד)

מס' תעודת הזהות של סטודנט/ית 2 (למלא עם מקף ('-') במקרה שההגשה ע"י סטודנט/ית אחד בלבד)

הדוא"ל של סטודנט/ית 2 (למלא עם מקף ('-') במקרה שההגשה ע"י סטודנט/ית אחד בלבד)

כתובת ה-github בו העלתם את הפרויקט (בחשבון של אחד הסטודנטים)

כתובת ה-youtube בו בהעלתם את סרטון ההסבר

בטופס זה ישנם שדות אותם חובה עליכם למלא והם מסומנים ב

המשך

הצגת בעיית הלמידה

נתון corpus (מאגר) מתויג של סיפורים (labeled corpus, כלומר כולל target values).
לכל סיפור יש תגית המציינת את המגדר של כותב הסיפור:

- 'm' – עבור כותב (male)
- 'f' - עבור כותבת (female)

עליכם לבנות מודל סיווג שתפקידו לסווג את המגדר של פסקה סיפורית כנ"ל.

החומרים בהם יהיה מותר להשתמש

מותר להשתמש בכל חומר אותו למדנו הכולל
- python בסיסי
- המודולים: numpy, pandas, sickit learn (sklearn), ובמודול re עבור regular expressions.
- בנוסף מותר להשתמש במודולים ובכלי ניתוח טקסט המופיעים במודל

החומרים בהם אסור להשתמש

- אסור להשתמש בשום מודול נוסף מלבד אלו המוזכרים לעיל
- אסור להשתמש בשום קובץ חיצוני.
- אסור לצרף רשימות של מילים (כולל stop words) ולהשתמש בהם לסיווג.

הקבצים המצורפים למטלה (3 קבצים):

קובץ 1. Corpus מתויג – עבור ה-training

שם הקובץ: annotated_corpus_for_train.csv
קובץ csv של ה-corpus המתויג. מדובר בקובץ שמכיל train data, בצורה גולמית ושיש להפוך אותו ל- feature vectors כפי שלמדנו. הקובץ מכיל 2 עמודות:

- 'story' – עמודה המכילה פסקה סיפורית
- 'gender' – עמודה המכילה את המגדר של כותב/ת הפסקה

קובץ 2. Corpus לא מתויג -עבור סיווג ה-test

שם הקובץ: corpus_for_test.csv
קובץ csv נוסף, המכיל פסקאות סיפוריות. הקובץ מכיל דוגמאות חדשות אותם יש לסווג.
הוא מכיל את העמודות הבאות:

- 'story' – עמודה המכילה פסקה סיפורית
- 'test_example_id' –המסמן את דוגמת ה-test

קובץ 3. מחברת הגשה ריקה להגשת התרגיל

שם הקובץ: Assignment5-text-analysis.ipynb - המחברת שתריצו בה את הקוד, ההסברים, הניסויים והתוצאות

המחברת אינה מכילה כל קוד מחייב, מלבד המלצות לטעינה וכתיבת הפלט (אותו יש להגיש גם כן כקובץ נפרד).
את הקוד שלכם יש להגיש בקובץ זה

בדיקת איכות המודל

מה נחשב מודל איכותי?

המדד הנבחר להערכת איכות המודל הוא מדד f1.

תזכורת f1=2*(precision * recall)/(precision + recall)

המדד שנבחר להעריך את איכות המודל - Average-f1

f1_male – מחשבים את f1 (כפי שמוזכר לעיל), כאשר מחשיבים את הכותבים כמחלקה החיובית ואת הכותבות כמחלקה השלילית.

f1_female – מחשבים את f1 (כפי שמוזכר לעיל), כאשר מחשיבים את הכותבות כמחלקה החיובית ואת הכותבים כמחלקה השלילית.

<u>Average_f1</u> – יחושב כך: **Average_f1=(f1_male + f1_female)/2**

שימו לב – מדובר בעצם ב- macro average של מדד f1 (כפי שנלמד בהרצאה)

הגשות במטלה:

1. הגשת חובה - סרטון קצר של 2-3 דקות (לא יותר), בו אתם מציגים ומסבירים מה עשיתים ואת התוצאות (תצטרכו להעלות את הסרטון ל- YouTube).
2. הגשת חובה – Assignment5-text-analysis.ipynb - קובץ ה-jupyter notebook המצורף למטלה (ללא שינוי שמו), המכיל את כל הקוד בו השתמשתם לצורך אימון המודל, וסיווג הדוגמאות החדשות.
   - הקוד אמור להריץ את כל השלבים שישמשו לבניית המודל ולסיווג הדוגמאות ב-test.
     עליכם להשתמש בקובץ ה-csv המייצג corpus המתויג לאימון ולסווג את דוגמאות האימון המופיעים ב-corpus הלא מתוייג, כפי שמפורט בפסקה הבאה (הסברים על ניקוד המטלה).
   - שימו אתם צריכים קוד עובד, מקורי שלכם, שיעבוד גם בסביבה שלנו ויפיק את אותם תוצאות
   - יש ללוות את הקוד שלכם בהערות הסבר בגוף הקוד.
   - יש להעלות את הקובץ לפרויקט שיפתח בדף ה-github של אחד המשתתפים

הסברים על ניקוד המטלה

הגשה תקינה בסיסית – <u>עד 35 נקודות</u>

- הגשה תקינה – הגשת סרטון הסבר ופרוייקט ב-github, הכולל תהליך תקין (המתואר בסעיף זה)
- מבחינת הקוד, צריך לכלול קריאה של הקלט, וקטוריזציה של הטקסטים,
  תהליך של אימון ושערוך המודל עם validation-set, כולל כמובן התוצאות, עם ה- Average_f1.
- ישום נכון של כל הנ"ל על ה-test-set
- בנוסף, המחברת צריך לכלול את חיזוי 5 הדוגמאות הראשונות ו-5 הודגמאות האחרונות של ה-test.
- הכל צריך גם להיות מוסבר בסרטון וגם בהסברים קצרים המלווים את הקוד במחברת.

- **שערוך** המודל בעזרת **cross-validation** (ולא hold-out) – **<u>5 נקודות</u>**
- **איכות המודל - Average_f1 של <u>0.6 לפחות</u> על cross-validation – <u>10 נקודות</u>**
- **איכות מודל משופרת** – **Average_f1 על cross-validation <u>עד 10 נקודות</u>** (בצורה מדורגת)
- **מורכבות ה-flow** על שלביו השונים – **<u>עד 20 נקודות</u>**
  - תהליך ה- preprocessing, וקטוריזציה, שימוש ברכיבים נוספים, נירמול של הטקסט
    פרמטרים שונים של בתהליך (כולל גם אלגוריתם הלמידה)
- **מורכבות הניסויים** – **<u>עד 25 נקודות</u>** (מלוא הנקודות יינתנו לניסויים מוצלחים במיוחד)
  - ניסויים הכוללים את השלבים השונים ב-flow
  - השוואה של רכיבים שונים ב-preprocessing, וקטוריזציה, שימוש ברכיבים נוספים,
    נירמול של הטקסט, פרמטרים (והייפר-פרמטרים) שונים של בתהליך (כולל גם אלגוריתם הלמידה)
  - יש להציג בצורה ברורה את השפעת הניסוים על תוצאות של Average_f1
    (על cross-validation) בצורה השוואתית.
- **הגשה ביחידות – בונוס של <u>5 נקודות</u>**

  - לא ניתן לצבור יותר מ100% על המטלה

הסברים נוספים יתווספו, אם יהיה בכך צורך בהתאם לשאלות <u>בפורום שאלות ותשובות</u> (נא לעקוב)

בהצלחה לכולם :-)